

DOCUMENT RESUME

ED 189 143

TM 800 313

AUTHOR Shuller, Stephen M.
TITLE A Large-Scale Assessment of an Instructional Program to Improve Test-Wiseness in Elementary School Students.
PUB DATE Oct 79
NOTE 17p.
AVAILABLE FROM Educational Solutions, Inc., 80 Fifth Ave., New York, NY 10011 (Single copy free)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Instructional Materials; *Reading Achievement; *Reading Tests; Standardized Tests; *Test Coaching; *Test Wiseness
IDENTIFIERS Metropolitan Achievement Tests; *Mini Tests; New York (New York)

ABSTRACT

Test wiseness (TW) has been recognized as an important source of variance on standardized achievement tests, independent of content or random error. To help students develop the type of TW skills necessary for optimal performance on reading achievement tests (ie. time using, guessing strategies), a comprehensive TW skills instructional program called Mini Tests was developed. The program consists of two sets of one page worksheets, one set designed to introduce TW skills, the other to provide practice in using them on materials similar to standardized reading achievement test items. As a preliminary assessment of the effectiveness of Mini Tests, data available regarding their use in the New York City Public Schools (grades 2-9) in 1974-76 were analyzed. The results indicated that New York City Elementary Schools which purchased Mini Tests made statistically significant gains in the numbers of their students scoring on grade level over those schools not purchasing Mini Tests. (GSK)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED189143

Tm 800313

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

A Large-scale Assessment of an Instructional Program
to Improve Test-wiseness in Elementary School Students

Stephen M. Shuller
Educational Solutions, Inc.

October, 1979

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

S. Shuller

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

A LARGE-SCALE ASSESSMENT OF AN INSTRUCTIONAL PROGRAM
TO IMPROVE TEST-WISENESS IN ELEMENTARY SCHOOL STUDENTS

Most school systems in the United States rely on norm-referenced standardized tests of achievement for information on which to base important decisions such as the makeup of instructional groupings, placement in educational programs, promotion, and even graduation. Since many of these decisions are vital to students and to the success of the schools' educational programs, it is important that these tests measure each student's achievement as accurately as possible, minimizing the effects of other factors.

Testwiseness (TW) -- "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situations to receive a high score" (Millman, Bishop, and Ebel, 1965) -- has been recognised as an important source of variance on standardized achievement tests, independent of content or random error (Ebel, 1965; Thorndike, 1951). More measurement error probably results from too little TW than from too much (Ebel, 1972).

Although TW appears to increase developmentally, perhaps due to the cumulative effect of test-taking experiences, large individual differences in TW are prevalent at all ages (Moore, Schutz, and Baker, 1966). These differences have been found to be independent of IQ and cognitive ability (Diamond and Evans, 1972; Dunn and Goldstein, 1959).

Millman et al. (1965) developed a typology of TW that has guided much of the subsequent research in this area (Sarnacki, 1979). This typology divides TW into "elements independent of the test constructor or test purpose" -- general TW skills such as time-using, error avoidance, and guessing strategies, and deductive reasoning skill -- and test-dependent elements -- test-specific skills such as cue-using strategies or consideration of the specific intent of the test constructors.

Investigation of subjects' employment of cue-using strategies with pseudo-achievement tests containing questions about fictitious information was the

focus of most early TW research. Bajtelsmit (1977), Gibb (1964), Moore et al. (1966), Walstrom and Boersma (1968), and Woodley (1973) all succeeded in increasing students' test scores by explicit instruction in cue-using strategies.

More recently, investigators have begun to measure the effects of TW instruction on standardized achievement test performance, with subjects ranging in age from pre-schoolers to junior high school students (Callenbach, 1973; Eakens, Green, and Bushell, 1976; Gaines and Jongsma, 1974; Gross, 1976; Oakland, 1972; Slaughter, 1977). In each of these studies, instruction was provided in general TW skills such as time-using and guessing strategies. In addition, Callenbach (1973), Eakens et al. (1976), and Oakland (1972) provided instruction in familiarity with directions and item formats of the particular standardized achievement test each was using to assess the effects of this instruction. Statistically significant correlations between TW instruction and test score improvement were found in all five studies.

Familiarity with test directions and item formats are test or test-type specific skills not mentioned in Millman et al.'s typology which appear to be important for optimal performance on standardized achievement tests. It can be argued that these skills are part of the subject matter of the test, in which case variability on these skills is not part of measurement error. This line of thought raises questions of content validity, however, which are seldom addressed by the major producers of standardized achievement tests (Buros, 1977). None of the presently available studies separates the effect of familiarity with directions and formats on test performance from that of more general TW skills.

The test-specific TW skills included in "intent consideration strategy" in Millman et al.'s (1965) typology also appear to be important factors in

achievement test performance, but they have received little attention from researchers. On reading achievement tests in particular, many items have two or more answers which can be considered correct in a general sense, or no answer that is exactly correct. Picking the "best answer," as most test directions state, is in part a process of considering the intent of the test constructors.

To help students develop all of these types of TW skills necessary for optimal performance on reading achievement tests, a comprehensive TW skills instructional program called Mini Tests (Hinman, Mitchell, Shuller, and Swartz, 1975, 1976; Hinman, Mitchell, and Swartz, 1976; Hinman and Spence, 1972) was developed. This program was prepared by analysing, with respect to TW skills, the most widely-used tests of reading achievement (those included in the Anchor Test Study, 1975), taking into account error patterns typical of students whose actual reading ability exceeded that measured by these tests. The program consists of two sets of one page worksheets, one set designed to introduce TW skills, the other to provide practice in using them on materials similar to standardized reading achievement test items.

It is expected that students who are instructed in TW skills with the Mini Test program will be able to maximize these skills. For those students who begin the program with less than optimal TW skills and, at the same time, at least minimal reading skills, an increase in reading achievement test scores can be expected as a consequence of improved TW skills. Those beginning with poorer TW skills can be expected to make comparatively larger gains on their reading test scores.

As a preliminary assessment of the effectiveness of Mini Tests, data available regarding their use in the New York City Public Schools (NYCPS) in the 1974-75 and 1975-76 school years has been analysed. This analysis is limited by the nature of these data, so that only speculative conclusions are

possible. Still, it is hoped that this study will provide a basis for further work of an experimental nature.

Method

Subjects. The unit of analysis is the New York City Public Elementary School. Data from 621 schools are included, of which 138 purchased Mini Tests in 1974-75 and 483 did not.

Materials. Mini Test Core Kits (Hinman and Spence, 1972) were used. These kits parallel the format and item types of the reading subtests of the Metropolitan Achievement Test (MAT), 1971 Edition (Durost, Baxler, Wrightstone, Prescott, and Balow, 1971). The kits consist of Vocabulary and Passages Series paralleling MAT Test 1: Word Knowledge, and Test 2: Reading Comprehension, respectively.

Each Mini Test worksheet is short (one page containing 5 to 15 items), and usually can be completed in less than ten minutes. This format is designed to reduce tension and fatigue, while providing a means for overcoming test anxiety and improving accuracy and speed by introducing the students to the tests gradually, interspersed with feedback and analysis sessions.

In each series, Starter Tests, which introduce TW skills, are given first, followed by a sequence of Practice Tests arranged roughly in order of reading difficulty. Ten to 20 instructional periods (of 45 to 90 minutes each) during a four to eight week span are recommended.

A Teachers Guide accompanies each kit, and no special teacher training is required.

Dependent Measures. In the Spring of each year, the New York City Public Schools administer a standardized test of reading achievement to all pupils in

grades 2 through 9. In 1974, the MAT was administered; in 1975, the school system switched to the New York City Reading Test (NYCRT), a secure form of the Stanford Achievement Test (Madden, Gardner, Rudman, Karlson, and Merwin, 1972). The data available from each test represents the percentage of students in each school scoring on or above grade level (hereafter called the "school reading score"), according to the norms of the test given.

For the entire school system, the mean 1975 school reading score was 11.49% higher than the mean 1974 score. Such a large difference indicates that the norms for the NYCRT were probably substantially lower than those of the MAT. Since the school reading scores from 1974 and 1975 correlated .9366, it was assumed that the effect of the different norms on the school reading scores was approximately uniform. Therefore, the difference between the 1974 and 1975 school reading scores for each school (hereafter, the "school difference score") was used to compare among schools the change in school reading achievement from 1974 to 1975.

Procedure. Data regarding Mini Test use was limited to the number of kits purchased by each school. It was assumed that, given the large number of schools purchasing Mini Tests in 1974-75, this number could be used as a relative indication of how extensively Mini Tests were used by the various schools.

The schools were divided into five "treatment groups:" (1) those not purchasing Mini Tests in 1974-75; (2) those purchasing 1 or 2 kits; (3) those purchasing 3 - 5 kits; (4) those purchasing 6 - 8 kits; and (5) those purchasing more than 8 kits. The mean school difference scores of these five groups were compared to determine the effects of Mini Tests on school reading achievement.

Results

A one-way fixed-effects analysis of variance, school difference score by Mini Test purchase group, was performed. The mean school difference scores by group are presented in Table 1. Table 2 summarizes the analysis of variance,

Table 1. 1974-75 School Difference Score broken down by number of Mini Test Kits purchased in 1974-75.

<u>Group</u>	<u># of Kits Purchased</u>	<u>N</u>	<u>Change in Pct. on Grade Level, 74-75</u>	
			<u>Mean</u>	<u>S.D.</u>
1	0	483	11.15	6.49
2	1 - 2	27	11.44	4.64
3	3 - 5	80	12.92	6.57
4	6 - 8	19	13.67	6.30
5	9+	12	14.41	6.99
Total			11.53	6.47

Table 2. One-way analysis of variance, school difference score by Mini Test purchase.

	<u>SS</u>	<u>df</u>	<u>MS</u>	<u>F</u>
Among groups	410.9	4	102.7	2.48
Within groups	25547.9	616	41.5	$\alpha = .05$
Total				$\omega^2 = .0095$

5.14 .016

which was significant at the .05 level, $F(4,616) = 2.48$.

The purchase groups were then collapsed into those schools purchasing Mini Tests vs those not purchasing these materials in 1974-75. The results of this analysis, reported in Table 3, show that, on the average, those schools purchasing Mini Tests increased their percent on grade level about 1.7% more than those not purchasing these materials. This difference in the mean difference scores was statistically significant at the .005 level, $t(619) = 2.735$.

Table 3. Comparison of schools buying and not buying Mini Tests in 1974-75 on school difference score.

	<u>N</u>	<u>Mean</u>	<u>S.D.</u>
Schools buying	138	12.86	6.23
Schools not buying	483	11.15	6.49
Total	621	11.53	6.47

Pearson product-moment correlations were also computed in the following manner. To check against the possibility that there was a systematic difference in initial reading levels between schools subsequently purchasing or not purchasing Mini Tests, the correlation between the 1974 school reading scores and Mini Tests purchased was computed. There was no significant correlation either by binary choice (bought/not-bought, $r = -.029$) or by the actual number of kits purchased per school ($r = -.039$). To estimate the size of the effect of Mini Test use on the difference scores, the correlation between these two variables was computed. The result was a correlation of .123, which is statistically significant at the .005 level.

Discussion

The results show that New York City Elementary Schools which purchased Mini Tests in 1974-75 made statistically significant gains in the numbers of their students scoring on grade level over those schools not purchasing Mini Tests during the same period. This conclusion is consistent with the hypothesis that Mini Tests improve the TW skills of those students who use this material.

Due to the quasi-experimental nature of this study, there are some important limitations that should be taken into consideration in interpreting the results. First, some of the measures available, such as the Mini Tests sales data, were related to the needed information -- in this case, actual use of Mini Tests in the schools -- in unknown ways, requiring some inferences to link what was known to what was required. An effort was made to ground these inferences firmly, as discussed above.

Secondly, the schools self-selected which "treatment group" to join, on a basis unknown to the experimenters. Although subsequent correlation analysis showed that there was no difference in 1974 reading scores between schools purchasing and not purchasing Mini Tests in 1974-75, there might have been motivational differences, covarying treatments which could account for the post-treatment between-group differences. Although these possibilities cannot be ruled out, it should be noted that all New York City Public Schools are under intense pressure to do well on the city-wide reading tests, so that motivation and interest in improving instruction are widespread.

Finally, needing to use the school as the unit of analysis makes interpretation of the size of the effect of the treatment difficult. It is not clear what accounting for 1.5% of the variance in school reading scores means in terms of between subjects variance on the test per se. Experimental

studies of individual student performance are necessary to illumine this area.

Within these limitations, the results of the present study probably underestimate the size of the treatment effect. The Mini Test materials available in 1974-75 were focused specifically on preparation for tests similar to the MAT, and not, therefore, optimally suited to preparation for the NYCRT, which is substantially different in format and types of questions asked. Thus, the results reported here represent improvement mainly on general TW skills, and to a small extent, reading-test-specific skills. The subsequently developed Mini Test Supplementary Kits, which take into account many more of the NYCRT test-specific skills, should result in greater improvement of TW skills, resulting in larger differences in reading scores between those schools using and not using Mini Tests. This hypothesis is the subject of Study 2.

The number of Mini Test kits sold almost certainly overestimates the actual usage, i.e. there is little doubt that some kits were purchased and not used, or used incompletely or incorrectly. This overestimation would also tend to reduce the observed effectiveness of the materials.

Mini Tests are hypothesized to improve reading achievement test scores only as a byproduct of improved TW skills. Therefore, those students who already have good test-taking skills cannot be expected to improve their test scores to any considerable degree through exposure to the program. Thus, the inclusion of students with good test-taking skills (due to lack of a control for this variable) would serve to lessen even more the observed effect of the Mini Test treatment.

Study 2

Subjects. Once again, the New York City Public Elementary School is the unit of analysis. Data from 621 schools were included, 194 of which purchased Mini

Tests in 1975-76 and 427 did not.

Materials. In 1975-76, Mini Test Supplementary Kits (Hinman, Mitchell, Shuller and Swartz, 1975) became available in addition to the Core Kits used the previous year. The Supplementary Kits contain materials covering those formats and item types on the NYCRT, but not covered by the Core Kits. Thus, the combination of Core and Supplementary Kits provide TW instructional materials optimally suited to the NYCRT.

Procedure. Once again, data regarding Mini Test use was limited to the number of kits purchased by each school. Again, it was assumed that purchase data could be used as an indication of the extent to which the purchased kits were used. The situation was complicated, however, by the fact that schools which purchased Core Kits during 1974-75 could have used left-over materials during 1975-76. There was no way to measure this Mini Test usage with the available data.

To control as well as possible for previous Mini Test use, and to control for differences in pre-treatment school reading scores, partial correlation analysis and the analysis of covariance were used.

Results

Partial correlations were computed between 1976 school reading scores and the numbers of Mini Test kits purchased by the schools, controlling for 1974 and 1975 school reading scores and numbers of Mini Test kits purchased in 1974-75. Separate partial correlations were computed between 1976 school reading scores and 1975-76 purchase of Core Kits ($r = .1539$), Supplementary Kits ($r = .2292$) and both types of kits combined ($r = .2151$). All of these correlations are significant at the .001 level.

The analysis of covariance was performed using a multiple regression procedure (Kerlinger, 1976). The dependent variable was the 1976 school reading score. Covariates were the 1974 and 1975 school reading scores, and the number of Core Kits purchased during 1974-75. Independent variables were number of Core and Supplementary Kits purchased in 1975-76. The results of this analysis, $F(3,612) = 16.95$, showed statistical significance at the .001 level.

To obtain an approximate measure of the amount of school reading test variance accounted for by Mini Test use, analyses of variance were calculated separately with numbers of Core and Supplementary Kits purchased in 1975-76 as the respective independent variables, and the difference between 1975 and 1976 school reading scores as criterion variable. The analyses of variance for the Core Kits showed statistical significance at the .01 level ($F(3,617) = 4.65$); for the Supplementary Kits at the .001 level ($F(3,617) = 11.05$). Core Kits accounted for 2.90% of the variance in the school difference scores, and Supplementary Kits for 7.22%. Combined use of the two types of kits accounted for 6.99% of the variance on the school difference scores.

Discussion

Since the Supplementary Kits contain instructional materials covering the formats and item types on the NYCRT, these kits provide test-specific TW instruction that was not available in 1974-75. The finding that Supplementary Kit usage accounted for 7.22% of the difference score variance, as compared with 1.5% for the Core Kits in 1974-75 suggests that TW instruction in test-specific skills is important. That combined use of both types of kits accounted for slightly less variance than use of Supplementary Kits alone is probably an artifact due to use of Core Kits purchased the previous year. Thus some schools which in this study appear to have used only Supplementary

Kits actually used both types of kit in 1975-76, and Core Kits in 1974-75 as well.

Although all the limitations of the first study apply to Study 2 as well, the available data is consistent with the conclusion that Mini Tests are of significant value in preparing students to do their best on standardized tests of reading achievement. The results of both studies lend support to Thorndike's (1951) and Ebel's (1965) contention that TW is an important factor in standardized achievement test scores, consistent with the results of Eakens et al. (1976), Callenbach (1973), and Oakland (1972).

Mini Tests differ substantially from the instructional materials used in the studies cited above in that Starter Tests are included. The Starter Tests are teaching instruments that lead students to discover for themselves the essential insights or awarenesses necessary for testwiseness. In contrast, other TW programs provide only practice material combined with verbal instructions. For this reason, it is expected that Mini Tests are more effective than other existing TW programs.

Unfortunately, none of the studies cited above reports the amount of variance accounted for by their treatments, or any other measure of the size of the treatment effect. Although different units of analysis preclude comparison of the amount of variance accounted for between the present studies and those previously published, this information is very important in assessing the educational and psychological significance of the programs, and should be reported.

Experimental studies comparing two or more of the TW instructional skills programs now available would enable the relative merits of these programs to be tested, and also provide further information about the extent to which TW is a factor influencing achievement test scores.

As these materials become more widely used, opportunities for large-scale controlled experimental tests will become possible.

References

- Buros, O.K. Fifty years in testing: Some reminiscences, criticisms, and suggestions. Educational Researcher, 1977, 9-15.
- Bajtelsmit, J.W. Test-wiseness and systematic desensitization programs for increasing adult test-taking skills. Journal of Educational Measurement, 1977, 14, 335-341.
- Callenbach, C. The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second grade students. Journal of Educational Measurement, 1973, 10, 25-30.
- Diamond, J.J. and Evans, W.J. An investigation of the cognitive correlates of testwiseness. Journal of Educational Measurement, 1972, 9, 145-150.
- Dunn, T.F. and Goldstein, L.G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 1959, 19, 171-179.
- Durost, W.N., Bixler, H.H., Wrightstone, J.W., Prescott, G.A., and Balow, I.H., Metropolitan Achievement Tests. New York: Harcourt Brace Jovanovich, 1970, 1971.
- Eakens, D.J., Green, D.S., and Bushell, D. The effects of an instructional test-taking unit on achievement test scores. Journal of Educational Research, 1976, 67-71.
- Ebel, R. Measuring Educational Achievement. Englewood Cliffs, N.J: Prentice-Hall, 1965.
- Ebel, R. Essentials of Educational Measurement (2nd Edition). Englewood Cliffs, N.J: Prentice-Hall, 1972.
- Gaines, W.G. and Jongsma, E.A. The effect of training in test-taking skills on the achievement scores of fifth grade pupils. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, April 1974.
- Gibb, B.G. Test-wiseness as a secondary cue response. (Doctoral Dissertation, Stanford University). Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.
- Gross, L.J. The effects of three selected aspects of test-wiseness on the standardized test performance of eighth grade students. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, April, 1976.
- Hinman, D.H., Mitchell, K.A., Shuller, S.M., and Swartz, T. Mini Test Supplementary Kits. New York: Educational Solutions, 1975.
- Hinman, D.H., Mitchell, K.A., Shuller, S.M., and Swartz, T. Mini Test Vocabulary Comprehension Kits. New York: Educational Solutions, 1976.

- Hinman, D.H., Mitchell, K.A., and Swartz, T. Mini Test Picture Reading Skills Kits. New York: Educational Solutions, 1976.
- Hinman, D.H. and Spence, I. Mini Test Core Kit. New York: Educational Solutions, 1972.
- Kerlinger, F. Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, 1968.
- Madden, R., Gardner, E.F., Rudman, H.C., Karlsen, B., and Merwin, J.C. Stanford Achievement Test. New York: Harcourt Brace Jovanovich, 1972.
- Millman, J., Bishop, C.H., and Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.
- Moore, J.C., Schutz, R.E., and Baker, R.L. The application of a self-instructional technique to develop a test-taking strategy. American Educational Research Journal, 1966, 3, 13-17.
- Oakland, T. The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. Journal of School Psychology, 1972, 10, 355-360.
- Sarnacki, R.E. An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 1979, 49, 252-279.
- Slaughter, B.A. Improving student performance through teaching test-taking skills. In A.J. Nitko (Ed.), Exploring Alternatives to Current Standardized Tests. Proceedings of the 1976 National Testing Conference, University of Pittsburgh, 1977.
- Thorndike, E.L. Reliability. In E.F. Lindquist (Ed.), Educational Measurement. Washington, D.C: American Council on Education, 1951.
- Wahlstrom, M. and Boersma, F.J. The influence of testwiseness upon achievement. Educational and Psychological Measurement, 1968, 28, 413-420.
- Woodley, K.K. Test-wiseness program development and evaluation. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, 1973.